

巻頭論文

データの価値を創出する データマネジメント技術



森田 温*



撫中達司**



郡 光則***

Data Management Technology that Creates Value of the Data

Atsushi Morita, Tatsuji Munaka, Mitsunori Kori

要 旨

ストレージ技術やネットワーク技術の進歩を背景に、データ量が急増しつつある。2007年に世界中で1年間に生み出されたデータの総量は281エクサバイトと言われている(1エクサバイト=2⁶⁰バイト)が、2011年には1,800エクサバイトに達すると予想されている。

また、データの内容も多様化している。従来の情報システムで主流であった構造化データに加え、文書、画像など、いわゆる非構造化データの増加が著しい。さらに近年では、ログやセンサデータに代表される、更新を伴わない追記型のデータが増加しつつある。

これら多様なデータから真に必要な情報を抽出し活用したいという要求にこたえるため、三菱電機では“データマネジメント技術”の研究開発に取り組んでいる。

当社のデータマネジメント技術は、様々なデータの高速処理を行う“高性能並列情報検索技術”を基盤として、

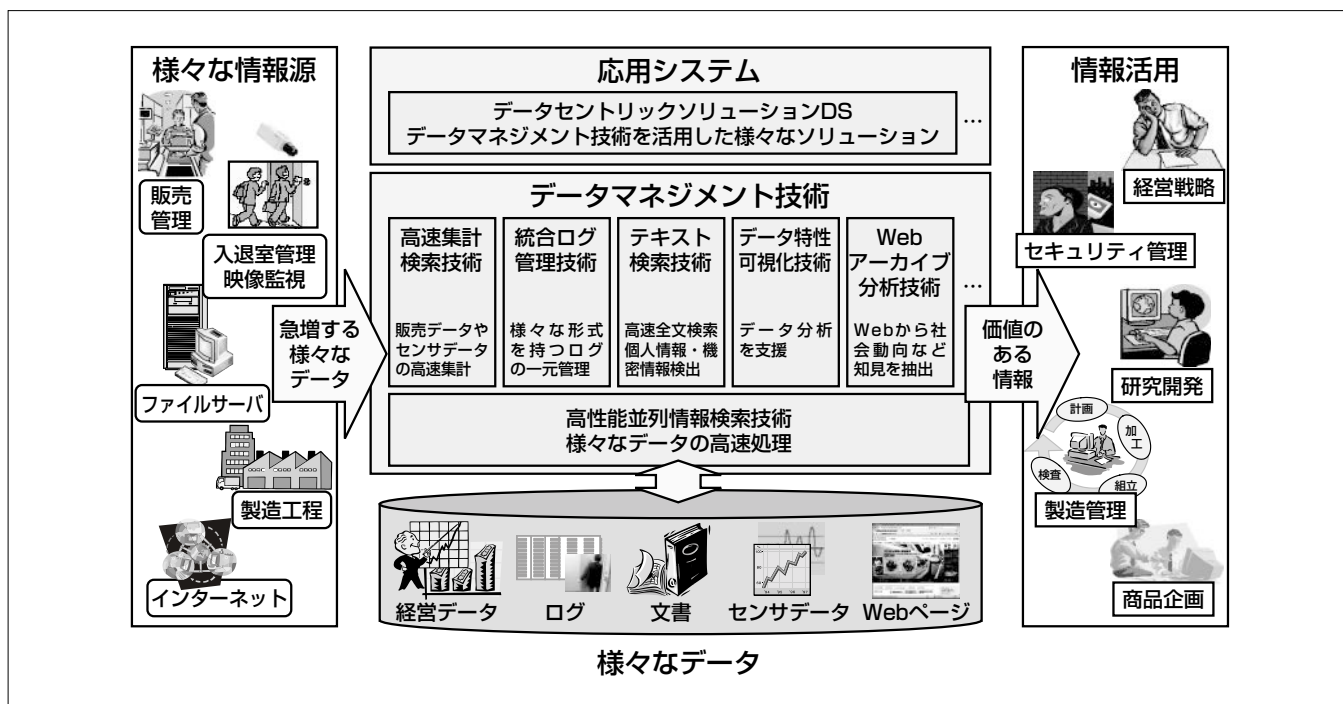
- (1) 販売データやセンサデータの集計検索を高速に行う高速集計検索技術

- (2) 様々な形式のログを一元的に管理し、イベント間の関係を抽出・管理する統合ログ管理技術
- (3) 文書の高速全文検索や個人情報・機密情報を高精度で抽出するテキスト検索技術
- (4) データの分析を支援するデータ特性可視化技術
- (5) 膨大なWebページから社会動向などの知見を抽出するWebアーカイブ分析技術

などの技術からなる。また、三菱電機インフォメーションテクノロジー(株)(MDIT)から、これらの技術を活用したソリューション“データセントリックソリューション^(注1)DS”が提供されている。

データマネジメント技術は、ビジネスインテリジェンス、文書管理・検索、ログ管理、センサデータ管理、Web情報の分析など、様々な分野で多くのユーザーに活用されている。

(注1) データセントリックソリューションは、三菱電機インフォメーションテクノロジー(株)の登録商標である。



三菱電機のデータマネジメント技術

当社は、急増する多様なデータから新たな知見を得て価値を創出する、“データマネジメント技術”の研究開発に取り組んでいる。データマネジメント技術は、高速処理の基盤技術“高性能並列情報検索技術”を基礎とし、各種データを活用するための“高速集計検索”“統合ログ管理”“テキスト検索”“データ特性可視化”“Webアーカイブ分析”などの技術からなる。MDITは、これらの技術を活用した“データセントリックソリューションDS”を提供しており、多くのユーザーに利用されている。

1. ま え が き

ストレージ技術やネットワーク技術の進歩を背景に、大量のデジタル化されたデータが生み出されている。このため、膨大な量の多種多様なデータの中から、真に必要な情報を抽出する“データマネジメント技術”の必要性が高まっている。

本稿では、当社の取り組んでいるデータマネジメント技術と、その応用事例について述べる。

2. データマネジメント技術の動向

2.1 データの量的、質的变化

ある調査⁽¹⁾によると、2007年に世界中で1年間に生み出されたデータの総量は281エクサバイトと言われている（1エクサバイト＝ 2^{60} バイト、約100万テラバイト）が、今後もデータ量は年率60%の比率で増加し続け、2011年には1,800エクサバイトに達すると予想されている（図1）。このため、膨大なデータの中から真に必要な情報を抽出し、活用することが困難になりつつある。

一方、データの内容も多様化している。従来の情報システムで扱われるデータの多くは、規則的な構造を持ついわゆる構造化データであったが、文書、画像、映像、音声、メールなど、明確な規則構造を持たない非構造化データ、XML(eXtensible Markup Language)やWebページのような半構造化データの増加が著しい。さらに近年では法規制に対応するための証跡データ（ログ）や、計測データ（センサデータ）などの、更新を伴わない追記型のデータが増加する傾向にある。このため、多種多様なデータの中に埋もれた情報の抽出と、異なる種類のデータ間の関係把握が課題となっている。

2.2 プラットフォームの動向

データマネジメント技術を支える情報処理プラットフォームの技術も変化しつつある。

2000年代前半まで続いたプロセッサの動作周波数の向上に代わり、性能向上の手段はプロセッサ（コア）数の増加を利用した並列処理に移行しつつある。このため、データ量増大に追従する性能向上を図るためには、並列処理の利用が不可欠となっている。

また、ディスク装置の容量は2000年以降年率60%の比率で増加し続けているが、その性能バランスは大きく変化している⁽³⁾。ディスク装置からメモリへのデータ転送速度は年率40%程度で向上しているが、機械的な動作を伴うアクセス（シーク、回転待ち）性能は年率16%程度の向上にとどまっている（図2）。このため、細かな単位でのランダムアクセスよりも、一度に大量のデータを読み書きした方が性能上有利になる傾向が拡大しつつある。データのアクセス効率を向上するため、データの一部を主メモリに格納する

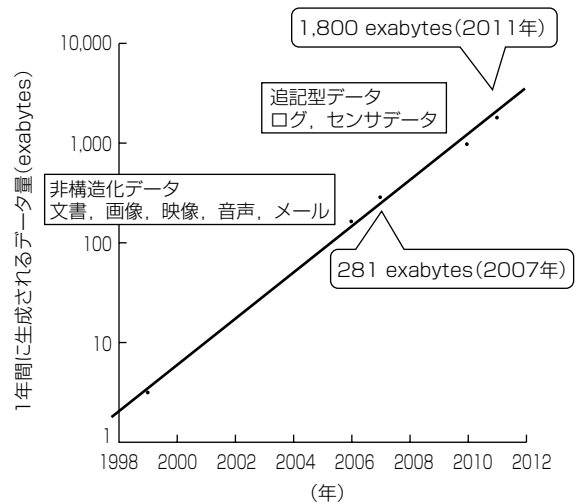


図1. データ量の爆発的増大⁽¹⁾⁽²⁾

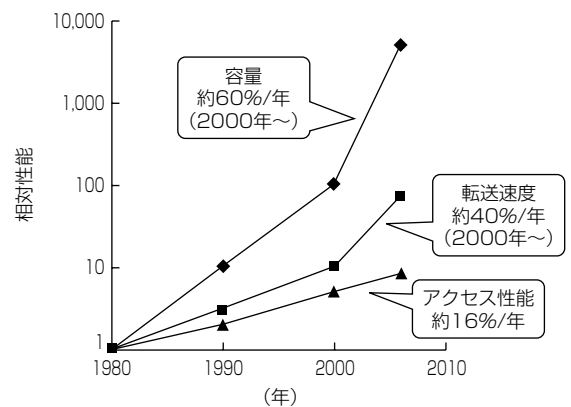


図2. ディスク装置の性能トレンド

キャッシュの技術が広く利用されている。しかし、この方法では対象データが増加してメモリに収まらなくなると急速に性能が低下する。急増するデータの効率的なアクセスには、ストレージアクセス自体の高速化が必要になる。

3. 当社のデータマネジメント技術への取り組み

当社は、爆発的に増加する多様なデータから真に必要な情報を抽出するため、様々なデータを効率的に管理し分析する技術、情報を安全に取り扱うための情報セキュリティの技術、映像・画像や音声などのマルチメディアデータを取り扱う技術などの研究開発に取り組んでいる。ここでは、データを管理し分析する技術“データマネジメント技術”を中心に、当社の取り組みについて述べる。

3.1 データマネジメント技術

当社のデータマネジメント技術は、高速処理を実現するための“高性能並列情報検索技術”を基礎とし、各種データを活用するための“高速集計検索”“統合ログ管理”“テキスト検索”“データ特性可視化”“Webアーカイブ分析”などの技術からなる（図3）。

3.1.1 高性能並列情報検索技術

高性能並列情報検索技術は、時間とともにデータが追加

される追記型データベースに適した高速処理技術であり、次の基本技術から構成される⁽⁴⁾。

- (1) プロセッサやディスク装置の利用効率を高め、切れ目なくデータを流すデータ駆動型の並列処理技術
- (2) 効率的データ配置と高速・高圧縮率のデータ圧縮によるストレージアクセス技術
- (3) 複雑な検索条件を高速処理する文字列照合方式と、学習型フィルタによるテキストフィルタリング技術
- (4) 様々な種類のデータを一元的に取り扱い、データの間

の関係を管理する異種データ統合管理技術

高性能並列情報検索技術は、近年のプラットフォーム動向を反映して並列処理の技術⁽¹⁾とストレージアクセスを効率化する技術⁽²⁾によって、データ量の増大に対応した高速処理を実現する。また、データから情報を抽出する技術⁽³⁾とデータの関係を管理するための技術⁽⁴⁾によって、多様化するデータの活用を図る。

3.1.2 高速集計検索技術

高速集計検索技術は、高性能並列情報検索技術を利用して、ビジネスインテリジェンスや計測データ管理でデータの集計や検索を高速に行う技術である⁽⁵⁾。

3.1.3 統合ログ管理技術

統合ログ管理技術は、セキュリティ管理や内部統制が必要とされる証跡(ログ)の一元的な管理を実現し、ログに含まれるイベントの間の関係を管理する技術である⁽⁶⁾。

3.1.4 テキスト検索技術

テキスト検索技術は、文書管理・検索でメールや文書、Webページなどをキーワードによって高速に検索し、また、個人情報や機密情報などの特定の意味を持った情報を高精度で抽出する技術である⁽⁷⁾。

3.1.5 データ特性可視化技術

データ特性可視化技術は、計測データ管理等で対話的な分析操作を効率化する技術である⁽⁸⁾。

3.1.6 Webアーカイブ分析技術

Webアーカイブ分析技術は、100億を超える膨大なWebページを分析して社会動向などの知見を抽出し、Web情報のより高度な活用を図るための技術である⁽⁹⁾。

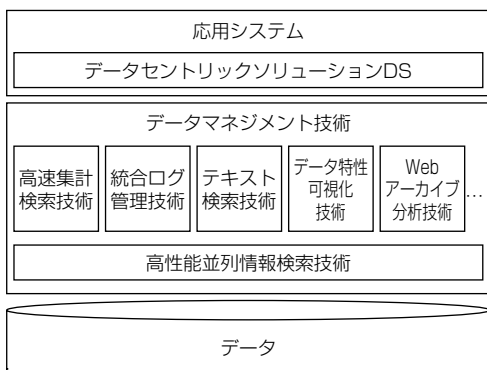


図3. 当社のデータマネジメント技術

3.2 データマネジメント技術の応用システム

データセントリックソリューションDSは、データマネジメント技術をエンドユーザーが容易に利用できるように製品化したMDITのソリューションであり、企業に蓄積されたデータの連携・統合・活用を効果的に実現する⁽¹⁰⁾。

4. データマネジメント技術の応用例

次に、当社のデータマネジメント技術の応用例について述べる。

4.1 ビジネスインテリジェンス

1990年代以降、それまで基幹系システムに使用されてきたデータベースとは別に、意思決定支援を目的とした情報分析向けのデータウェアハウスと呼ばれるデータベースが出現し、流通業における販売データ分析などで利用されるようになった。

データウェアハウスを用いたビジネスインテリジェンスと呼ばれるデータ分析では、例えば大量の売上げ明細データを日付、地域、商品のような複数の分析軸(視点)ごとに集計してデータの傾向を把握する。このような分析手法はOLAP(On-Line Analytical Processing)と呼ばれており、分析軸ごとの集計処理(多次元集計処理)の高速実行が必要になる。

当社では、高性能並列情報検索技術と高速集計検索技術による高速集計検索エンジンを開発した。このエンジンはMDITの“データ分析プラットフォームDIAPRISM”⁽¹¹⁾に組み込まれて販売されている(図4)。

4.2 文書管理・検索

電子化された文書の急増を背景に、1990年代から文書に人手でキーワードを付加せずに自由なキーワードで検索する全文検索エンジンが普及した。

全文検索では、事前に転置(inverted)索引を作成し、索引を用いて検索することによって高速化を図る場合が多い。しかし、全文検索用の索引のサイズは元の文書よりも大きくなることも多く、大規模文書の全文検索では索引の読み出しの効率化が重要な課題となる。

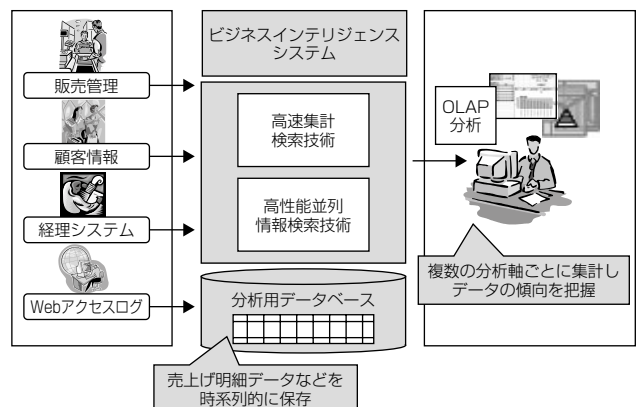


図4. ビジネスインテリジェンスシステム

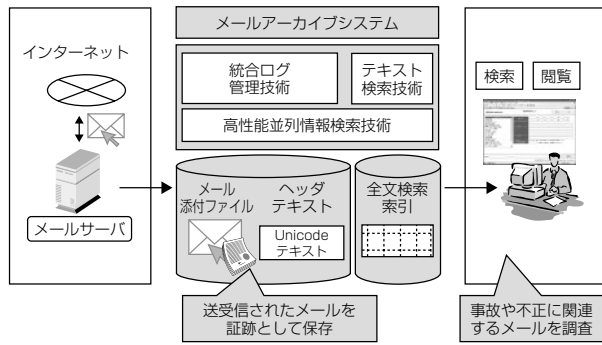


図5. メールアーカイブシステム

当社では、高性能並列情報検索技術を適用した高性能全文検索エンジンを開発した⁽¹²⁾。1テラバイトの文書を1秒で検索できる高速検索や、Unicodeによる多言語に対応した検索を特長とする。この全文検索エンジンは、特許検索システムやメールアーカイブシステム“LogAuditor^(注2) Mail Saver”⁽¹³⁾ (図5)などで活用されている。

また近年、人手で単純に指定できない複雑な条件による検索の必要性が高まっている。例えばメールアーカイブシステムで、外部へ送信されたメールに個人情報や機密情報が含まれていないか確認しようとしても、その検索条件の作成は容易ではない。また、複雑な検索条件を作成できても、検索に時間がかかるという課題があった。

当社では、複雑な検索条件にでも事前に索引を作成することなく1億文字/秒の高速照合が可能な高速文字列照合技術を開発した。この技術は三菱スペース・ソフトウェア株(MSS)の個人情報検出ツール“すみずみ君^(注3)”，メールフィルタ“LeakStopper^(注3) for SMTP(Simple Mail Transfer Protocol)”に組み込まれて販売されている。

さらに学習型フィルタを併用することによって、高精度、高速に機密情報を検出する技術を開発した⁽¹⁴⁾。学習型フィルタはあらかじめ用意したサンプルデータから出現頻度の高い部分文字列を統計的特徴として学習し、新たな入力データと統計的特徴の類似性を比較することによって検索を行う。このため、人手で複雑な検索条件を作成する必要がない。機密文書を用いた評価の結果、99.9%を検出する高い検出精度を確認した。

4.3 ログ管理

インターネットや電子化データの普及と相まって、情報漏洩(ろうえい)事故がしばしば発生するようになり、企業におけるセキュリティ管理の重要性に対する認識が高まっている。また、いわゆる日本版SOX法(金融商品取引法)の施行を契機に、企業内の内部統制の強化の一環として情報システムの生成するログを長期間保存することが求められるようになった。これに対応して、ログ管理向けの専用

(注2) LogAuditorは、三菱電機インフォメーションテクノロジー株の登録商標である。

(注3) すみずみ君, LeakStopperは、三菱スペース・ソフトウェア株の登録商標である。

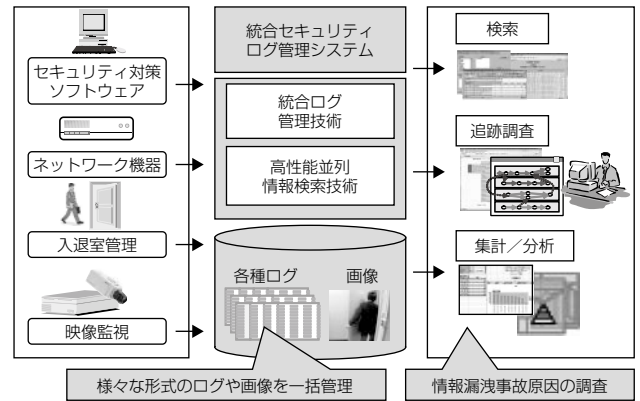


図6. 統合セキュリティログ管理システム

データベースが出現するようになった。

ログは1日あたり数億件が蓄積される事例もあり、そのデータ量は数十テラバイトにも及ぶ場合がある。このため、高速な蓄積・検索やストレージコスト削減が課題となる。また、ログの形式は情報システムごとに様々であるが、従来のデータベースでは格納時に形式を統一する必要があるが、統合管理が困難であった。さらに近年では、監視映像や送受信されたメールなども証跡として保存・管理したいという要求があるが、RDBMS(Relational Data Base Management System)に代表される従来の汎用(はんよう)データベースでは、これら様々なデータを効率的に管理することが困難であった。

当社では高性能並列情報検索技術を応用し、様々なデータを一元管理するログデータベースを開発した。ログデータベースはいかなる形式のデータも格納可能な“本文”と、本文に付随する“属性”の組によってデータを管理することによって、各種ログ、メール、文書、画像を含む、あらゆる形式のデータを一元管理できる。

一例として、ログデータベースを入退室管理システムや映像監視システムと連携した統合セキュリティログ管理システムを示す(図6)。入室した時点の画像を自動的に統合ログデータベースに取得し、入退室ログと関連付けて検索できる。さらにデータの間の関係を管理し、ログの内容把握をより容易にする履歴追跡型ログデータベースに、特定の人や物に関連する行動を詳細に追跡できる。

ログデータベースはMDITの“LogAuditor Enterprise”^{(15) (16)}でログや監視画像の格納に利用されているほか、“LogAuditor Mail Saver”でメールや添付ファイルの保存に利用されている。

4.4 センサデータ管理

近年、多数のセンサによって計測されたデータを、ネットワークを通じて収集するセンサデータベースの構築が進みつつある。例えば地球温暖化対策の一環として、改正省エネ法(エネルギーの使用の合理化に関する法律)では事業者エネルギー使用量の把握を義務付けており、電力消費

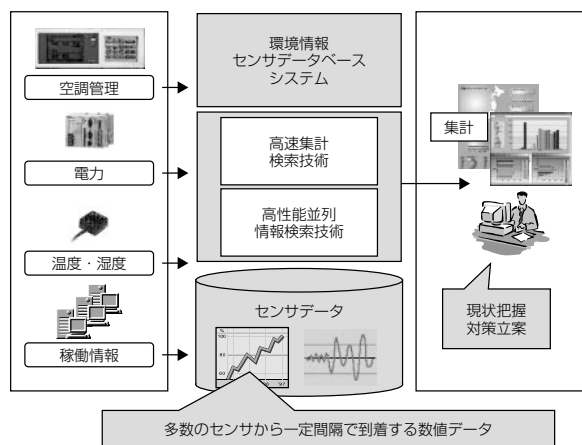


図7. 環境情報センサデータベースシステム

量の収集・管理が必要となっている。

このような要求に対応するため、MDITでは高性能集計検索エンジンを適用した環境経営推進ソリューション“MELGREEN”⁽¹⁷⁾を販売している。

また当社では、高速集計検索エンジンをセンサデータ管理に適用したセンサデータベースの研究開発を進めている。センサデータベースでは、多数のセンサから一定間隔で到着するデータの実時間的な反映、同一時間帯に到着したデータの一括処理、数値データ向けの効率的なデータ圧縮などが課題であり、これらを解決する技術の開発に取り組んでいる(図7)。

4.5 Web情報の分析

Webは、実世界の様々な事象を即時に把握するには不可欠のメディアとなっている。当社は、産学連携プロジェクト等を通じ、120億ページ分の情報を蓄積した大規模日本語Webアーカイブを構築し、Web情報の分析技術の研究開発を行ってきた。

複数のWebページを内容の関連性に基づいてグループ化し、グループ相互のつながり(情報構造)を視覚化するとともに、過去から現在に至る情報の発展・伝播(でんぱ)過程を提示し、マーケティングリサーチや社会動向の把握、注目話題分析など、Web情報のより高度な活用を目指している。

5. む す び

データマネジメント技術を取り巻く動向と当社の取り組みについて述べた。データ量の増加と多様化は当分継続するものと予想されており、データから必要な情報を抽出し、価値を創出するデータマネジメント技術の重要性は今後ますます高まるものと考えている。今後もデータマネジメント技術の研究開発への取り組みを続けていく。

参 考 文 献

(1) The Expanding Digital Universe, IDC White Paper

(2008)

- (2) Lyman, P., et al.: How Much Information? 2003
<http://www2.sims.berkeley.edu/research/projects/how-much-info-2003>
- (3) Gray, J.: Long Term Storage Trends and You
http://research.microsoft.com/en-us/um/people/gray/talks/io_talk_2006.ppt
- (4) 郡 光則, ほか: 高性能並列情報検索技術, 三菱電機技報, **83**, No.12, 705~708 (2009)
- (5) 山岸義徳, ほか: 高速集計検索エンジンとセンサデータベースへの応用, 三菱電機技報, **83**, No.12, 709~712 (2009)
- (6) 平井規郎, ほか: 統合ログ管理技術と履歴追跡型ログデータベース, 三菱電機技報, **83**, No.12, 713~716 (2009)
- (7) 加藤 守, ほか: テキスト検索技術と個人情報・機密情報検出, 三菱電機技報, **83**, No.12, 717~720 (2009)
- (8) 藤野友也, ほか: データ特性の対話的可視化技術, 三菱電機技報, **83**, No.12, 721~724 (2009)
- (9) 田村孝之, ほか: 大規模日本語Webアーカイブの構築とその分析, 三菱電機技報, **83**, No.12, 725~728 (2009)
- (10) 荻野 正, ほか: 企業データをトータルに活用するデータセントリックソリューションDS, 三菱電機技報, **83**, No.12, 729~732 (2009)
- (11) 安藤隆朗, ほか: データウェアハウスソリューションとその適用効果の実際, 三菱電機技報, **77**, No.4, 267~270 (2003)
- (12) 中村隆顕, ほか: 大規模並列全文検索エンジンにおける多言語検索対応索引方式, 第8回情報科学技術フォーラム講演論文集(2), 115~116 (2009)
- (13) 大塚哲史, ほか: 1000万件のメールを1秒で検索する“LogAuditor Mail Saver”, 三菱電機技報, **82**, No.7, 461~464 (2008)
- (14) 機密情報検出ソフトウェアを開発, 三菱電機ニュースリリース (2009)
<http://www.mitsubishielectric.co.jp/news/2009/0325.htm>
- (15) 郡 光則, ほか: 多種多様なログの統合管理を実現する“LogAuditor Enterprise”, 三菱電機技報, **80**, No.10, 615~618 (2006)
- (16) 小山明伸, ほか: 物理セキュリティの統合管理を実現した“LogAuditor”, 三菱電機技報, **83**, No.9, 563~566 (2009)
- (17) 松井陽子, ほか: 省エネルギーのPDCAの管理基盤環境経営推進ソリューション“MELGREEN”, 三菱電機技報, **83**, No.7, 413~416 (2009)